## GLOWPICK - PRODUCT INFORMATION ##

Sys.setlocale("LC\_COLLATE", "ko\_KR.UTF-8")

Sys.setlocale("LC\_CTYPE", "ko\_KR.UTF-8")

Sys.setlocale("LC\_MONETARY", "ko\_KR.UTF-8")

Sys.setlocale("LC\_NUMERIC", "ko\_KR.UTF-8")

Sys.setlocale("LC\_TIME", "ko\_KR.UTF-8")

Sys.getlocale()

#Sys.setenv(JAVA\_HOME = '/Library/Java/JavaVirtualMachines/jdk1.8.0\_201.jdk/Contents/Home/jre')

#dyn.load('/Library/Java/JavaVirtualMachines/jdk1.8.0\_201.jdk/Contents/Home/jre/lib/server/libjvm.dylib')

#Sys.getenv("JAVA\_HOME")

#remove.packages("rJava")

#install.packages("rJava")

#install.packages('DBI')

#install.packages('RJDBC')

library(rvest)

library(rJava)

library(DBI)

library(RJDBC)

perfume.review <- NULL # set variable to save data from crawling

all\_pages<-NULL

for(i in 3325:3375){

#26223270~3300

url <-paste0("https://www.glowpick.com/product/",i)

text <- read\_html(url, encoding="UTF-8")

# a) brandName

nodes <- html\_nodes(text, ".product-main-info\_\_brand\_name")

brandName <- html\_text(nodes)

brandName

# b) productName

nodes <- html\_nodes(text, ".product-main-info\_\_product\_name")

productName <- html\_text(nodes)

productName <- gsub("\n", "", productName)

productName <- gsub(pattern = "(^ +| +$)", replacement = "", productName)

productName

# c) volume/price

nodes <- html\_nodes(text, ".product-main-info\_\_volume\_price")

volumePrice <- html\_text(nodes)

volumePrice <- gsub("\n", "", volumePrice)

volumePrice <- gsub(pattern = "(^ +| +$)", replacement = "", volumePrice)

test <- strsplit(volumePrice, "/")

result <- unlist(test)

volume <- result[1]

price <- result[2]

price <- gsub("원","",price)

price <- gsub(",","",price)

price

volume

price <- as.numeric(price)

price

# d) rating

nodes <- html\_nodes(text, ".ratings\_\_score")

rating <- html\_text(nodes)

rating <- gsub(pattern = "(^ +| +$)", replacement = "", rating)

rating <- as.numeric(rating)

rating

# e) category

nodes <- html\_nodes(text, ".product-detail\_\_category")

category <- html\_text(nodes)

category <- gsub("\n", "", category)

category <- gsub(pattern = "(^ +| +$)", replacement = "", category)

category

category <- strsplit(category, " ")

category[1]

category[2]

# f) description

nodes <- html\_nodes(text, ".product-detail\_\_description")

desFull <- html\_text(nodes)

desFull <- gsub("\n", "", desFull)

desFull <- gsub(pattern = "(^ +| +$)", replacement = "", desFull)

test <- substr(desFull,1,70)

result <- unlist(test)

des <- result[1]

des

# g) tags

nodes <- html\_nodes(text, ".product-detail\_\_tags")

tags <- html\_text(nodes)

tags <- gsub(pattern = "(^ +| +$)", replacement = "", tags)

tags

if(length(tags)==0){

tags <- "NONE"

tags

} else {

tags

}

# h) images

library(httr)

h <- read\_html(url)

imgs <- html\_nodes(h, 'img')

imgsrcList <- html\_attr(imgs, 'src')

imgsrcAdr <- imgsrcList[3]

imgsrcAdr <- gsub("https://d9vmi5fxk1gsw.cloudfront.net/home/glowmee/upload/", "", imgsrcAdr)

imgsrcAdrList <- strsplit(imgsrcAdr, "/")

imgsrcAdrList <- unlist(imgsrcAdrList)

image <- imgsrcAdrList[2]

image

keywords<-"keyword"

counts<-0

# h) image download

#download.file(imgsrcList[3], image)

if(length(category)>1){

pagetest1 <- data.frame(p\_category=category[1], br\_name=brandName,

p\_name=productName, p\_volume=volume, p\_price=price, p\_image=image,

p\_keyword=keywords, p\_des=des, p\_count=counts, t\_name=tags, stringsAsFactors = FALSE)

pagetest1

colnames(pagetest1)

colnames(pagetest1)<-c('p\_category','br\_name','p\_name','p\_volume','p\_price','p\_image','p\_keyword','p\_des','p\_count','t\_name')

pagetest1

pagetest2 <- data.frame(p\_category=category[2], br\_name=brandName,

p\_name=productName, p\_volume=volume, p\_price=price, p\_image=image,

p\_keyword=keywords, p\_des=des, p\_count=counts, t\_name=tags, stringsAsFactors = FALSE)

pagetest2

colnames(pagetest2)

colnames(pagetest2)<-c('p\_category','br\_name','p\_name','p\_volume','p\_price','p\_image','p\_keyword','p\_des','p\_count','t\_name')

pagetest2

page<-rbind(pagetest1,pagetest2)

page

} else {

pagetest1 <- data.frame(p\_category=category[1], br\_name=brandName,

p\_name=productName, p\_volume=volume, p\_price=price, p\_image=image,

p\_keyword=keywords, p\_des=des, p\_count=counts, t\_name=tags, stringsAsFactors = FALSE)

pagetest1

colnames(pagetest1)

colnames(pagetest1)<-c('p\_category','br\_name','p\_name','p\_volume','p\_price','p\_image','p\_keyword','p\_des','p\_count','t\_name')

page<-pagetest1

page

}

# marge all data

all\_pages<-rbind(all\_pages,page)

}

for(i in 3270:3300){

#26223270~3300

url <-paste0("https://www.glowpick.com/product/",i)

text <- read\_html(url, encoding="UTF-8")

# a) brandName

nodes <- html\_nodes(text, ".product-main-info\_\_brand\_name")

brandName <- html\_text(nodes)

brandName

# b) productName

nodes <- html\_nodes(text, ".product-main-info\_\_product\_name")

productName <- html\_text(nodes)

productName <- gsub("\n", "", productName)

productName <- gsub(pattern = "(^ +| +$)", replacement = "", productName)

productName

# c) volume/price

nodes <- html\_nodes(text, ".product-main-info\_\_volume\_price")

volumePrice <- html\_text(nodes)

volumePrice <- gsub("\n", "", volumePrice)

volumePrice <- gsub(pattern = "(^ +| +$)", replacement = "", volumePrice)

test <- strsplit(volumePrice, "/")

result <- unlist(test)

volume <- result[1]

price <- result[2]

price <- gsub("원","",price)

price <- gsub(",","",price)

price

volume

price <- as.numeric(price)

price

# d) rating

nodes <- html\_nodes(text, ".ratings\_\_score")

rating <- html\_text(nodes)

rating <- gsub(pattern = "(^ +| +$)", replacement = "", rating)

rating <- as.numeric(rating)

rating

# e) category

nodes <- html\_nodes(text, ".product-detail\_\_category")

category <- html\_text(nodes)

category <- gsub("\n", "", category)

category <- gsub(pattern = "(^ +| +$)", replacement = "", category)

category

category <- strsplit(category, " ")

category[1]

category[2]

# f) description

nodes <- html\_nodes(text, ".product-detail\_\_description")

desFull <- html\_text(nodes)

desFull <- gsub("\n", "", desFull)

desFull <- gsub(pattern = "(^ +| +$)", replacement = "", desFull)

test <- substr(desFull,1,70)

result <- unlist(test)

des <- result[1]

des

# g) tags

nodes <- html\_nodes(text, ".product-detail\_\_tags")

tags <- html\_text(nodes)

tags <- gsub(pattern = "(^ +| +$)", replacement = "", tags)

tags

if(length(tags)==0){

tags <- "NONE"

tags

} else {

tags

}

# h) images

library(httr)

h <- read\_html(url)

imgs <- html\_nodes(h, 'img')

imgsrcList <- html\_attr(imgs, 'src')

imgsrcAdr <- imgsrcList[3]

imgsrcAdr <- gsub("https://d9vmi5fxk1gsw.cloudfront.net/home/glowmee/upload/", "", imgsrcAdr)

imgsrcAdrList <- strsplit(imgsrcAdr, "/")

imgsrcAdrList <- unlist(imgsrcAdrList)

image <- imgsrcAdrList[2]

image

keywords<-"keyword"

counts<-0

# h) image download

#download.file(imgsrcList[3], image)

if(length(category)>1){

pagetest1 <- data.frame(p\_category=category[1], br\_name=brandName,

p\_name=productName, p\_volume=volume, p\_price=price, p\_image=image,

p\_keyword=keywords, p\_des=des, p\_count=counts, t\_name=tags, stringsAsFactors = FALSE)

pagetest1

colnames(pagetest1)

colnames(pagetest1)<-c('p\_category','br\_name','p\_name','p\_volume','p\_price','p\_image','p\_keyword','p\_des','p\_count','t\_name')

pagetest1

pagetest2 <- data.frame(p\_category=category[2], br\_name=brandName,

p\_name=productName, p\_volume=volume, p\_price=price, p\_image=image,

p\_keyword=keywords, p\_des=des, p\_count=counts, t\_name=tags, stringsAsFactors = FALSE)

pagetest2

colnames(pagetest2)

colnames(pagetest2)<-c('p\_category','br\_name','p\_name','p\_volume','p\_price','p\_image','p\_keyword','p\_des','p\_count','t\_name')

pagetest2

page<-rbind(pagetest1,pagetest2)

page

} else {

pagetest1 <- data.frame(p\_category=category[1], br\_name=brandName,

p\_name=productName, p\_volume=volume, p\_price=price, p\_image=image,

p\_keyword=keywords, p\_des=des, p\_count=counts, t\_name=tags, stringsAsFactors = FALSE)

pagetest1

colnames(pagetest1)

colnames(pagetest1)<-c('p\_category','br\_name','p\_name','p\_volume','p\_price','p\_image','p\_keyword','p\_des','p\_count','t\_name')

page<-pagetest1

page

}

# marge all data

all\_pages<-rbind(all\_pages,page)

}

for(i in 8900:8930){

#2622

url <-paste0("https://www.glowpick.com/product/",i)

text <- read\_html(url, encoding="UTF-8")

# a) brandName

nodes <- html\_nodes(text, ".product-main-info\_\_brand\_name")

brandName <- html\_text(nodes)

brandName

# b) productName

nodes <- html\_nodes(text, ".product-main-info\_\_product\_name")

productName <- html\_text(nodes)

productName <- gsub("\n", "", productName)

productName <- gsub(pattern = "(^ +| +$)", replacement = "", productName)

productName

# c) volume/price

nodes <- html\_nodes(text, ".product-main-info\_\_volume\_price")

volumePrice <- html\_text(nodes)

volumePrice <- gsub("\n", "", volumePrice)

volumePrice <- gsub(pattern = "(^ +| +$)", replacement = "", volumePrice)

test <- strsplit(volumePrice, "/")

result <- unlist(test)

volume <- result[1]

price <- result[2]

price <- gsub("원","",price)

price <- gsub(",","",price)

price

volume

price <- as.numeric(price)

price

# d) rating

nodes <- html\_nodes(text, ".ratings\_\_score")

rating <- html\_text(nodes)

rating <- gsub(pattern = "(^ +| +$)", replacement = "", rating)

rating <- as.numeric(rating)

rating

# e) category

nodes <- html\_nodes(text, ".product-detail\_\_category")

category <- html\_text(nodes)

category <- gsub("\n", "", category)

category <- gsub(pattern = "(^ +| +$)", replacement = "", category)

category

category <- strsplit(category, " ")

category[1]

category[2]

# f) description

nodes <- html\_nodes(text, ".product-detail\_\_description")

desFull <- html\_text(nodes)

desFull <- gsub("\n", "", desFull)

desFull <- gsub(pattern = "(^ +| +$)", replacement = "", desFull)

test <- substr(desFull,1,70)

result <- unlist(test)

des <- result[1]

des

# g) tags

nodes <- html\_nodes(text, ".product-detail\_\_tags")

tags <- html\_text(nodes)

tags <- gsub(pattern = "(^ +| +$)", replacement = "", tags)

tags

if(length(tags)==0){

tags <- "NONE"

tags

} else {

tags

}

# h) images

library(httr)

h <- read\_html(url)

imgs <- html\_nodes(h, 'img')

imgsrcList <- html\_attr(imgs, 'src')

imgsrcAdr <- imgsrcList[3]

imgsrcAdr <- gsub("https://d9vmi5fxk1gsw.cloudfront.net/home/glowmee/upload/", "", imgsrcAdr)

imgsrcAdrList <- strsplit(imgsrcAdr, "/")

imgsrcAdrList <- unlist(imgsrcAdrList)

image <- imgsrcAdrList[2]

image

keywords<-"keyword"

counts<-0

# h) image download

#download.file(imgsrcList[3], image)

if(length(category)>1){

pagetest1 <- data.frame(p\_category=category[1], br\_name=brandName,

p\_name=productName, p\_volume=volume, p\_price=price, p\_image=image,

p\_keyword=keywords, p\_des=des, p\_count=counts, t\_name=tags, stringsAsFactors = FALSE)

pagetest1

colnames(pagetest1)

colnames(pagetest1)<-c('p\_category','br\_name','p\_name','p\_volume','p\_price','p\_image','p\_keyword','p\_des','p\_count','t\_name')

pagetest1

pagetest2 <- data.frame(p\_category=category[2], br\_name=brandName,

p\_name=productName, p\_volume=volume, p\_price=price, p\_image=image,

p\_keyword=keywords, p\_des=des, p\_count=counts, t\_name=tags, stringsAsFactors = FALSE)

pagetest2

colnames(pagetest2)

colnames(pagetest2)<-c('p\_category','br\_name','p\_name','p\_volume','p\_price','p\_image','p\_keyword','p\_des','p\_count','t\_name')

pagetest2

page<-rbind(pagetest1,pagetest2)

page

} else {

pagetest1 <- data.frame(p\_category=category[1], br\_name=brandName,

p\_name=productName, p\_volume=volume, p\_price=price, p\_image=image,

p\_keyword=keywords, p\_des=des, p\_count=counts, t\_name=tags, stringsAsFactors = FALSE)

pagetest1

colnames(pagetest1)

colnames(pagetest1)<-c('p\_category','br\_name','p\_name','p\_volume','p\_price','p\_image','p\_keyword','p\_des','p\_count','t\_name')

page<-pagetest1

page

}

# marge all data

all\_pages<-rbind(all\_pages,page)

}

all\_pages

# connect JDBC

drv\_Oracle <- JDBC(driverClass="oracle.jdbc.driver.OracleDriver", classPath="/Users/jenkim/R\_Workspace/ojdbc6.jar")

con\_Oracle <- dbConnect(drv\_Oracle, "jdbc:oracle:thin:@//182.237.126.19:1521/xe","javabig20","1234")

dbWriteTable(con\_Oracle,"Product",all\_pages,overwrite=TRUE)

# DB disconnect

dbDisconnect(con\_Oracle)

# delete temp objects

rm(drv, con\_Oracle, query)

# file download

write.csv(perfume.review, "perfume.csv")

write.table(perfume.review, "perfume.txt")

library(openxlsx)

write.xlsx(all\_pages, "perfume.xlsx", sheetName="perfume1")

#########################################################review

library(RSelenium)

library(dplyr)

library(xlsx)

remDr<-remoteDriver(remoteServerAddr="localhost",port=4444,browserName="chrome")

remDr$open()

remDr$navigate(urls)

all.name<-c()

all.age <-c()

all.gender <-c()

all.score<-c()

all.review<-c()

all.index<-c()

#3325~3375, 3270~3300, 8900~8930

for(i in 8900:8930){

urls<-paste0(url<-'https://www.glowpick.com/product/',i)

remDr$navigate(urls)

Sys.sleep(1)

all.index<-c(all.index,i)

webElem <- remDr$findElement("css selector", ".section-list-item")

webElem$findElement(using = "css selector", ".section-list-item")$clickElement()

webElem <- remDr$findElement("css selector", "body")

webElem$sendKeysToElement(list(key = "end"))

Sys.sleep(0.2)

webElem$sendKeysToElement(list(key = "end"))

Sys.sleep(0.2)

webElem$sendKeysToElement(list(key = "end"))

Sys.sleep(1)

webElem$sendKeysToElement(list(key = "end"))

Sys.sleep(1)

webElem$sendKeysToElement(list(key = "end"))

webElem <- remDr$findElement("css selector", "body")

Sys.sleep(1)

#나이

agedom<-remDr$findElements(using="css selector","#gp-product-detail .info .txt ")

agedom

age <-sapply(agedom,function(x){x$getElementText()})

age<-unlist(age)

age<-substr(age,start=1,stop = 3)

all.age <-c(all.age,age)

Sys.sleep(1)

#성별

genderdom<-remDr$findElements(using="css selector","#gp-product-detail .info .icon-gender-f,#gp-product-detail .info .icon-gender-m")

genderdom

gender<-sapply(genderdom,function(x){

x$getElementAttribute('class')

}

)

gender<-substr( gender,start=25,stop = 26)

gender<- gsub("f", "여성",gender)

gender <- gsub("m", "남성",gender)

all.gender <-c(all.gender,gender)

Sys.sleep(1)

#평점

scoredom<-remDr$findElements(using="css selector","#gp-product-detail .info .icon-sprite.rating-grade-icon.gpa-good-small,#gp-product-detail .info .icon-sprite.rating-grade-icon.gpa-best-small,#gp-product-detail .info .icon-sprite.rating-grade-icon.gpa-soso-small,#gp-product-detail .info .icon-sprite.rating-grade-icon.gpa-bad-small,#gp-product-detail .info .icon-sprite.rating-grade-icon.gpa-worst-small")

scoredom

score<-sapply(scoredom,function(x){

x$getElementAttribute('class')

}

)

score <-substr(score ,start=35,stop = 38)

all.score<-c(all.score,score)

#리뷰

reviewdom<-remDr$findElements(using="css selector",".review")

review<-sapply(reviewdom,function(x){x$getElementText()})

review <- gsub("\t", "", review)

review <- gsub("\r\n", "", review)

review <- gsub("\n", "", review)

all.review<-c(all.review,review)

#향수이름 리뷰 개수만큼 반복해서 cbind

ss<-length(review)

namedom<-remDr$findElements(using="css selector",".product-main-info\_\_product\_name\_\_text")

name <-sapply(namedom,function(x){x$getElementText()})

name<-unlist(name)

name

all.name

z<-1

while(z<=ss) {

all.name <-c(all.name,name)

z<-z+1

}

ss<-1

length(all.name)

length(all.review)

}

reviewf3<-cbind(all.name,all.age,all.gender,all.score,all.review)

##################

colnames(reviewf3)=c("이름","나이","성별","평점","리뷰")

data<-data.frame(reviewf3)

# file download

write.csv(reviewf3, "perfume.cvs")

write.xlsx(reviewf3,"perfume.xlsx")